Utilizing Tighter ELBO Analysis for Semi-Supervised Multi-Label Learning with Deep VAE Models

Haozhe Feng

State Key Lab of CAD& CG, Real Doctor AI Research Centre

fenghz@zju.edu.cn

June 17, 2019

Outline

- Related Work
- Our Method and Results
- Further Work Plan

Related Work

- Regularization Based Deep Semi-Supervised Method
- Generation Method

Our Method and Results

Utilizing VAE to deal with semi-supervised learning tasks has been proposed in 2014[1] and 2017 [2]. They all use maximum likelihood estimation as the loss function, that is, maximize $\log p(x)$ while unsupervised and maximize $\log p(x, y)$ while supervised. The ELBO loss can be written as

$$\log p_\theta(x, y) \geq E_{q_{phi}(z|x,y)}[\log p_\theta(x|y, z) + \log p_\theta(y) + \log p(z) - \log q_\phi(z|x, y)] \tag{1}$$

$$= -L_l(x, y) \tag{2}$$

$$logp_\theta(x) \geq E_{q_{phi}(y,z|x)}[\log p_\theta(x|y, z) + \log p_\theta(y) + \log p(z) - \log q_\phi(y, z|x)] \tag{3}$$

$$= \sum_y q_\phi(y|x)(-L_l(x, y)) + H_{q_\phi(y|x)} \tag{4}$$

$$= -L_u(x) \tag{5}$$

Our Method and Results

However, the traditional loss functions $(1) - (5)$ have three drawbacks:

1. $-L_u(x)$ needs to compute $-L_l(x, y)$ many times
2. In real implementation, this model needs 2 stage hierarchical structure, which is unstable and hard to train
3. The loss function can't lead to a natural cross-entropy classification loss form, which is essential to the classification performance.
4. The multi-label form of $L_u(x)$ is

$$\sum_{y^1} \sum \cdots \sum_{y^k} q_\phi(y^1, \ldots, y^k|x)(L_l(x, y^1, \ldots, y^k)) - H_{q_\phi(y|x)} \tag{6}$$

and the complexity is $O(k^c)$

Our Method and Results

To deal with these 3 problems, we propose a new framework for the VAE based semi-supervised learning tasks with tigher ELBO analysis.
Firstly, instead of the assumption that (x, c) are pairs from the original data space, we view label c as the discrete latent variables obey multinominal distribution then make the independent assumption for both marginal and conditional distribution as following

$$p(z, c) = p(z)p(c) \tag{7}$$
$$p(z, c|x) = p(z|x)p(c|x) \tag{8}$$
$$q_\phi(z, c|x) = q_\phi(z|x)q_\phi(c|x) \tag{9}$$

Our Method and Results

Then utilizing the decomposition of KL divergence for independent random variable

$$\mathcal{D}_{KL}(p(z,c|x)\|q_\phi(z,c|x)) = \mathcal{D}_{KL}(p(z|x)\|q_\phi(z|x)) + \mathcal{D}_{KL}(p(c|x)\|q_\phi(c|x)) \tag{10}$$

we can rewrite the ELBO as

$$\log p(x) - \mathcal{D}_{KL}(q_\phi(z|x)\|p(z|x)) - \mathcal{D}_{KL}(q_\phi(c|x)\|p(c|x)) = E_{(z,c)\sim q_\phi(z,c|x)} \log p_\theta(x|z,c) - \mathcal{D}_{KL}(q_\phi(z|x)\|p(z)) - \mathcal{D}_{KL}(q_\phi(c|x)\|p(c)) \tag{11}$$

which can be used to train the unlabeled data and split their latent space into the continuous and discrete latent variables.

Our Method and Results

For the labeled data, we can get a similar loss function form of(11) by **Jensen's inequality**

$$\log p(x) = \log E_{z\sim q_\phi(z|x),c\sim p(c|x)} \frac{p(x,z,c)}{q_\phi(z|x)p(c|x)} \geq E_{z\sim q_\phi(z|x),c\sim p(c|x)} \log \frac{p(x,z,c)}{q_\phi(z|x)p(c|x)}$$

$$= E_{z\sim q_\phi(z|x),c\sim p(c|x)} \log p(x|z,c) + E_{z\sim q_\phi(z|x),c\sim p(c|x)} \log \frac{p(z)p(c)}{q_\phi(z|x)p(c|x)}$$

$$= E_{z\sim q_\phi(z|x),c\sim p(c|x)} \log p(x|z,c) - \mathcal{D}_{KL}(q_\phi(z|x)\|p(z)) - \mathcal{D}_{KL}(p(c|x)\|q_\phi(c|x))$$

$$+ E_{c\sim p(c|x)} \log \frac{p(c)}{q_\phi(c|x)}$$

$$\approx^{\text{when} q_\phi(c|x)\approx p(c|x)} E_{z\sim q_\phi(z|x),c\sim p(c|x)} \log p(x|z,c) - \mathcal{D}_{KL}(q_\phi(z|x)\|p(z)) - \mathcal{D}_{KL}(p(c|x)\|q_\phi(c|x))$$

$$+ E_{c\sim q_\phi(c|x)} \log \frac{p(c)}{q_\phi(c|x)} \quad (12)$$

$$= E_{z\sim q_\phi(z|x),c\sim p(c|x)} \log p_\theta(x|z,c) - \mathcal{D}_{KL}(q_\phi(z|x)\|p(z)) - \mathcal{D}_{KL}(q_\phi(c|x)\|p(c))$$

$$- \mathcal{D}_{KL}(p(c|x)\|q_\phi(c|x)) \quad (13)$$

Our Method and Results

To make $\mathcal{D}_{KL}(q_\phi(z|x)\|p(z)), \mathcal{D}_{KL}(q_\phi(c|x)\|p(c))$ more explainable, we utilize the mutual information decomposition form[3]

$$E_{\hat{p}(x)}[\mathcal{D}_{KL}(q_\phi(z|x)\|p(z)] = I_{q_\phi}(x;z) + \mathcal{D}_{KL}(q_\phi(z)\|p(z)) \tag{14}$$

We can view $I_{q_\phi}(x;z)$ as a constant C_z. For $\hat{p}(x) = \frac{\sum_{i=1}^{N}\delta_{x_i}(x)}{N}$, we can rewrite (14) as the individual form

$$\sum_{i=1}^{N}\frac{\mathcal{D}_{KL}(q_\phi(z|x_i)\|p(z) - C_z}{N} = \mathcal{D}_{KL}(q_\phi(z)\|p(z)) \tag{15}$$

Our Method and Results

We utilize (15) to rewrite (13) as the final loss function form for the labeled data.

$$\begin{aligned}\arg_{\phi,\theta}\min \sum_{i=1}^{N} \frac{1}{N} E_{z\sim q_\phi(z|x_i), c\sim p(c|x_i)} \log -p_\theta(x_i|z,c) \\ - |\mathcal{D}_{KL}(q_\phi(z|x_i)\|p(z)) - C_z| - |\mathcal{D}_{KL}(q_\phi(c|x_i)\|p(c)) - C_c| \\ - \mathcal{D}_{KL}(p(c|x_i)\|q_\phi(c|x_i)) \end{aligned} \quad (16)$$

Our Method and Results

The new semi-supervised loss function form (13) has 3 benefits:

1. The cross entropy loss $\mathcal{D}_{KL}(p(c|x)\|q_\phi(c|x))$ is naturally derived
2. The framework can extend to multi-label task with complexity $O(kc)$
3. The one-stage training process is more stable than two-stage
4. By training with (13), the VAE model can reach a tighter ELBO
   To reach the optimial point, we need to minimize $\mathcal{D}_{KL}(p(c|x)\|q_\phi(c|x))$, which minimize$\mathcal{D}_{KL}(q_\phi(c|x))\|p(c|x)$ at the same time, and make the marginal equation (11) between $\log p(x)$ and ELBO tighter.

Our Method and Results

We have tried our method in the MNIST,SVHN and CheXpert dataset. Following are some results

(a) MNIST with 100 labeled data

M1+TSVM	M2	M1+M2	TighterVAE
11.82 (± 0.25)	11.97 (±1.71)	3.33 (±0.14)	3.31 (±0.19)

(b) SVHN with 4000 labeled data

M1+KNN	M1+TSVM	M1+M2	TighterVAE
65.63 (± 0.15)	54.33 (± 0.11)	36.02 (± 0.10)	31.92 (± 0.14)

(c) CheXpert with 5 pathologies

病理	Baseline 100%	Baseline 50%	TighterVAE 50%	TighterVAE 10%
Atelectasis	0.808	0.803	0.826	0.781
Cardiomegaly	0.834	0.821	0.855	0.786
Consolidation	0.904	0.897	0.891	0.816
Edema	0.898	0.892	0.9	0.868
Pleural Effusion	0.921	0.904	0.904	0.883


Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling.
Semi-supervised learning with deep generative models.
In *Advances in neural information processing systems*, pages 3581–3589, 2014.


Siddharth Narayanaswamy, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr.
Learning disentangled representations with semi-supervised deep generative models.
In *Advances in Neural Information Processing Systems*, pages 5925–5935, 2017.


Shengjia Zhao, Jiaming Song, and Stefano Ermon.
Infovae: Information maximizing variational autoencoders.
arXiv preprint arXiv:1706.02262, 2017.